Contributor(s):

- **Amália Mendes**, Centre for Linguistics at the University of Lisbon (CLUL), amalia.mendes@clul.ul.pt

Reference Corpus of Contemporary Portuguese

The Reference Corpus of Contemporary Portuguese (CRPC) is a large electronic Portuguese corpus that has been under development at the Centre for Linguistics of the University of Lisbon (CLUL) since 1988. Presently, this corpus contains 311.4 million words from written and spoken language, reflecting both regional and national varieties of Portuguese (Portugal, Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, São Tomé and Principe, Goa, Macao and East Timor) and covering different text types (e.g., newspapers, books, periodicals, parliament sessions, decisions of the Supreme Court of Justice, leaflets, correspondence, miscellaneous; informal and formal recording situations). The corpus is lemmatized and tagged with POS, and the written subpart can be searched online through the CQPWeb. Several lexical applications have been developed based on the CRPC, such as: a frequency lexicon of European Portuguese with POS and lemma information, a comparative description of the vocabulary of the African varieties of Portuguese, and a lexicon of multi word expressions. The CRPC has also been a source of data for the Dictionary of Contemporary Portuguese of the Lisbon Academy of Sciences (published in 2001), providing example sentences and helping to establish the list of entries. The size and diversity of the CRPC makes it an important resource for Portuguese e-lexicography.

Lexicon of Multi Word Expressions (LEX-MWE-PT)

A lexicon of multi word expressions (MWE) extracted from a 50M words subcorpus of CRPC. The selection of the MWE relied on the MI statistical measure, raw frequency and manual validation of lexical and syntactic fixedness and total or partial loss of compositional meaning. In all, the lexicon comprises 14,153 group lemmas and 48,154 word combinations and provides corpus contexts for each MWE. A version is available on Meta-Share.

Frequency Lexicon of European Portuguese

A frequency lexicon extracted from a 16 M words subcorpus of CRPC, with POS and lemma information.
http://www.clul.ul.pt/sectores/linguistica_de_corpus/lmcpc/lmcpc_dec.zip

Comparative Lexicon of the African varieties of Portuguese

A corpus-based contrastive lexicon of the lemmas (and word forms) occurring in each variety, with frequency data and genre distribution. http://www.clul.ul.pt/en/research-teams/87-linguistic-resources-for-the-study-of-the-african-varieties-of-portuguese-r

Lexicographic Corpus of Portuguese

Corpus of lexicographic works (from the XVI to the XVIII centuries) in digitised version, available for online concordance queries (DICIweb). It comprises 23 texts and a total of 7 million tokens. http://clp.dlc.ua.pt

Contributor(s):

- **Lars Trap-Jensen**, Society For Danish Language and Literature, Denmark, ltj@dsl.dk
- **Henrik Lorentzen**, Society For Danish Language and Literature, Denmark, hl@dsl.dk

Ordnet.dk website

The website *ordnet.dk* consists of three main parts: The Danish Dictionary (Den Danske Ordbog, DDO), Dictionary of the Danish Language (Ordbog over det danske Sprog, ODS) and KorpusDK. Data from the two dictionaries, both originally print dictionaries, one historical and one modern, are brought together with a contemporary reference corpus and a wordnet, all with Danish as the object language. Innovative aspects include data exploitation across the components, such as onomasiological queries in the dictionary based on thesaurus and wordnet data, and cross resource look-up possibilities from the three components. The modern DDO also contains pronunciation as sound files.

The Danish Dictionary contains almost 100,000 entries and 120,000 meaning descriptions and is regularly updated with new and revised entries. The Dictionary of the Danish Language is now a historical dictionary, first published in 28 volumes between 1918 and 1956, with five supplementary volumes published 1992-2005. In total, ODS contains about 225,000 entries and is the most comprehensive monolingual dictionary of Danish published so far.

In addition to these basic components, *ordnet.dk* provides access to digitized historical dictionaries: *Holbergordbogen*, describing in 35,000 entries the author Ludvig Holberg's complete works, and *Moths Ordbog*, the earliest monolingual Danish dictionary compiled c. 1700 but never published until the digitized version appeared in 2013. The dictionary contains 110,000 entries. These dictionaries are not (yet) integrated with the primary dictionaries.

All four dictionaries are also available as apps for iOS and Android.

Contributor(s):

- **Pascale Renders**, University of Liège, France, pascale.renders@ulg.ac.be
- **Yan Greub**, ATILF, France, Yan.Greub@atilf.fr

The e-FEW project

ATILF (Nancy, France) & ULg (Liège, Belgium)

Our project aims at providing researchers with a free, new and more efficient tool for using the 25 volumes of the *Französisches Etymologisches Wörterbuch* (FEW), written since 1920 by von Wartburg *et al*. This reference dictionary of Romance linguistics contains all the words of Galloromance languages and dialects from the 9th to the 20th century. The fact that it is currently underused results from the complexity of its structures. Each article gathers under its etymon-lemma words belonging to the same lexical family. Lexical units are classified according to etymological, phonetical, morphological, semantical, geographical and chronological criteria; wordclass, sense, chronological information and bibliographical references are systematically provided, but not always in an explicit way.

Updated articles are now directly published online and the 25 printed volumes are being digitalized. Their content will be automatically enriched with XML tags, allowing complex queries. This project is part of an international initiative, led by Nancy's ATILF laboratory with the collaboration of the University of Liège and the Trier Center for Digital Humanities.

The electronic FEW has four purposes. First, it aims at providing users with new ways of accessing the dictionary, including the possibility of multi-criteria queries. Secondly, it will facilitate the reading of articles, for example by providing hyperlinks between articles, by expanding abbreviations, etc. Thirdly, the electronic FEW aims at integrating the many additions and corrections produced by the scientific community in papers, monographies and other printed or electronic dictionaries (DEAF, AND, TLF-Etym, etc.). Finally, the interconnection between the FEW and these online dictionaries and other linguistics resources is another important goal of the project.

URL of the project: www.atilf.fr/few

Contributor(s):

- **Antton Gurrutxaga**, Elhuyar Foundation, Basque County, Spain, a.gurrutxaga@elhuyar.com
- **Klara Ceberio Berger**, Elhuyar Foundation, Basque County, Spain, k.ceberio@elhuyar.com
- **Igor Leturia**, Elhuyar Foundation, Basque County, Spain, i.leturia@elhuyar.com

e-Dictionaries at Elhuyar Foundation

Elhuyar Hiztegiak site (http://hiztegiak.elhuyar.org/)

Bilingual dictionary site, including three dictionaries:

Dictionary	Language	Entries / Subentries
Diccionario Elhuyar Hiztegia	eu-es/es-eu	91,000 / 23,000
Dictionnaire Elhuyar Hiztegia	eu-fr/fr-eu	36,600 / 3,800
Elhuyar Dictionary Hiztegia	eu-en/en-eu	28,600 / 3,300

The site has two areas:

- The free area, accessible to every user: dictionary consultation, last updates, popular entries; word of the day; different ways to interact with the Elhuyar team (new entries, corrections, proposals…); Blog
- The customer area (registration required):
  - A free dictionary for Android devices (eu <->es)
  - A free dictionary for Kindle eReader (eu -> es)
  - New content and services: additional notes; external links, favourite words, search history, grammar section; automatic retrieving from the Elhuyar es-eu parallel Web corpus (http://webcorpusak.elhuyar.org/cgi-bin/kontsulta2.py?mota=arrunta) of more examples of the entry's equivalents
  - Personalized language consultation service

In the current version, purchasers of Euskara-Gaztelania/Castellano-Vasco Elhuyar Hiztegia or the Euskara-Gaztelania/Castellano-Vasco Elhuyar Hiztegi Txikia 4. edizioa can register (registration code in the book). In 2014, the customer area will be enhanced and improved in order to make it accessible by subscription. The new services and contents will be based on NLP tools developed by Elhuyar I+G (http://www.elhuyar.org/hizkuntza-zerbitzuak/EN/R-D).

The Automatic Bilingual Dictionaries' site

Presentation: https://cordis.europa.eu/wire/index.cfm?fuseaction=article.Detail&rcn=37327

Website: http://hiztegiautomatikoak.elhuyar.org/bilaketa/en

Technical articles:

- X. Saralegi, I. Manterola, I. San Vicente. 2012. Building a Basque-Chinese Dictionary by using English as a Pivot: *In Proceedings of the 8th international conference on*

Language Resources and Evaluation, LREC'12. 23-25 May, Istanbul, Turkia.

- X. Saralegi, I. Manterola, I. San Vicente. 2011. Analizing Methods for Improving Precision of Pivot Based Bilingual Dictionaries: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, Edinburgo.

Elhuyar Encyclopaedic Dictionary of Science & Technology (ZTH)

Website (in Basque): http://zthiztegia.elhuyar.org/

Overview:

- 50 domains from three main areas: Mathematics, Natural Sciences and Technology
- 23,000 concepts (definitions in Basque)
- Terminology: eu, en, es, fr
- Encyclopaedic articles (607)
- Illustrations (1,500)
- 200 experts from different organizations (universities, institutes of technology, industrial companies…)

Future development:

- A new workbench for cooperative dictionary making (ongoing work, to be released in 2014). The core data structure will be a terminological DB
- Extending of the network of experts and collaborators. Special interest in the field of education (teachers and students)
- Quality assessment of content (expert network), terminology and language (Elhuyar Foundation)
- The contents will be released under a CC BY-SA license

Elhuyar Learner's Dictionary

Basque monolingual dictionary for learners. Overview of the current edition: http://www.elhuyar.org/hizkuntza-zerbitzuak/EN/Elhuyar-ikaslearen-hiztegia

In the short term (1-2 months), the dictionary will have a new website accessible to the purchasers of the printed edition.

Future development:

- Student dictionary (for students aged 12-18), with student/school-oriented macrostructure/microstructure
- Definitions in Basque & multilingual equivalences (es, en, fr)
- Corpus based
- Language in use: Corpus occurrences in context; didactic units: grammar links, exercises; related reading: articles, wikis; audio-visuals: illustrations, videos; Basque Academy links
- Interactivity: customization; feedback (community)
- Challenges and difficulties
    - Building a curriculum corpus (currently non-existent in Basque)
    - Integrating current monolingual and multilingual resources: sense distinctions

do not always match

- Advanced linguistic tools to exploit Basque corpora for (semi) automated lexicographic work (ongoing research work)

Elhuyar Dictionary of Idioms & Collocations

This dictionary project will use as main data source the ElhWebCorp corpus (http://webcorpusak.elhuyar.org/cgi-bin/kontsulta.py?mota=arrunta). The preliminary result of the extraction process can be freely consulted at the "Word Combinations" section of the Elhuyar Web Corpora site (http://webcorpusak.elhuyar.org/cgi-bin/kolokatuak.py).

Contributor(s):

- **Frits van der Kuip**, Fryske Akademy, the Netherlands, fvdkuip@fryske-akademy.nl
- **Hindrik Sijens**, Fryske Akademy, the Netherlands, hsijens@fryske-akademy.nl
- **Pieter Duijff**, Fryske Akademy, the Netherlands, pduijff@fryske-akademy.nl

Our project is an Online Dutch-Frisian dictionary. This new dictionary will replace a more concise and outdated paper dictionary, published in 1985. The new dictionary provides non-Frisian speaking people of a tool to write Frisian, but the main goal is to provide Frisian speaking people with a tool that helps them to write Frisian in a natural way, that is to say, not coloured by the dominant Dutch language. Most Frisian speakers have a lack of education in Frisian and are used to write Dutch, and to some extent they are illiterate in their mother tongue. The provincial government recognizes that the written use of Frisian can help to preserve the language. Therefore the project is supported by the provincial government.

To serve the users as optimal as possible the dictionary will not only give more entries than the 1985 dictionary, but will also pay much more attention to the application possibilities of Frisian by giving examples, collocations and idioms in plenty. Because the dictionary will be published on-line in stages, this will provide users to comment on the text and make suggestions to the editors (crowdsourcing).

At this moment we are preparing and planning the project. The provincial government provides funding for ICT-help and developping software. End of 2014 we hope to present a demo version of the on-line tool. The editing is planned for the next five years.

Contributor(s):

- **Ilga Migla**, Latvian Language Institute of the University of Latvia, ilga.migla@inbox.lv

The lexicographers from the Latvian Language Institute of the University of Latvia are working on a "Dictionary of the Modern Latvian Language" (DMLL) – an explanatory monolingual (Latvian) dictionary. The project ends on April, 2014. Then this dictionary will be accessible in the link www.tezaurs.lv/mlvv.

Currently the dictionary is usable in this link with glossaries which have a-o and r letters. The expected volume of DMLL is 50 000 – 60 000 glossaries. DMLL is a dictionary based on text corpus as well as other electronical resources which is written from anew and primarily is expected to be published on the internet in an electronical format.

Parallel to the internet resources when writing the DMLL, a special vocabulary file was made for this dictionary consisting of 160 000 units. DMLL is a dictionary that does not limit itself to analyse only some isolated part of the vocabulary (e.g. foreign words, terms) but also reflects the modern Latvian vocabulary as a whole, revealing both, the semantics of new words, and the semantic changes of the Latvian base vocabulary.

The dictionary includes various layers of the Latvian language; also some restrictions apply to the included widespread non-literal vocabulary, such as, Germanic barbarisms, etc. The dictionary offers new, politically neutral explanations to some financial, economical, political, philosophical, literary science and art science terms. Plant and animal explanations have been supplemented with corresponding family or kin (in several cases – species) in Latin.

DMLL also shows indications to loanwords about the origin of the word and also to the neologisms coined by the Latvian linguists, men of letters and cultural workers. At separate entries on a gray coloured background a commentary from the cultural point of view of the language has been added, if the corresponding entry is associated with mispronunciation, slips in orthography or troubles with the correct usage of grammatical forms, if a spelling change has taken place to the corresponding entry, as well as, if a word, word meaning or an expression into Latvian language is undesirable and it would be best to find a more suitable replacement.

Contributor(s):

- Prof. **Włodzimierz Gruszczyński**, The Institute of the Polish Language at the Polish Academy of Sciences, History of the Polish Language of the 17th and 18th Century Section, Poland, wlodekiewa@poczta.onet.pl

Electronic Dictionary of the Polish Language of the 17th and the 18th century (***e-SJP XVII***)

(Pol. *Elektroniczny słownik języka polskiego XVII i XVIII wieku*)

URL: http://xvii-wiek.ijp-pan.krakow.pl/pan_klient/ or http://sxvii.pl

This dictionary is the main project in progress being carried out by the Section leading by me. Now the dictionary is being published only on the Internet. The electronic version of the dictionary was initiated in 2004. Between 2004-2006 I prepared the project of electronic version – a very elaborate structure of a relational database – and I supervised the implementation (by M. Żółtak) as well as also develop the project.

A database was created. The entries are collected on the server and are being introduced by an electronic form. The authors of the entry articles have access to the dictionary in this way. They also have direct access to digitized source material: scans flashcards, source texts, bibliographical information, etc. For readers, a textual version of a transparent entry is generated with a possibility of being printed. *E-SJPXVII* contains currently about 16 thousand entries at various stages of the preparation of editorial content. The electronic version includes the highly modified contents of the 1st volume of *SJPXVII* printed on paper before 2004 and a large part of the reworked material of the *Jan Chryzostom Pasek's Language Dictionary*. New entries, beginning with B and C, are being successively added in electronic version, as well as entries beginning with other letters, because the change of the way of work and publication does not require an alphabetical order anymore.

Grammatical Dictionary of Polish **(*SGJP*) by Z. Saloni, M. Woliński, R. Wołosz, W. Gruszczyński, D. Skowrońska**

(Pol. *Słownik gramatyczny języka polskiego*)

URL: http://sgjp.pl/index.html.en

I am also co-author of *SGJP*, electronic dictionary on CD-ROM (more information see on the home page). We are now preparing on-line version of this dictionary.

Contributor(s):

- **Piotr Żmigrodzki**, Institute of Polish Language, Polish Academy of Science, Krakow, piotr@ijp-pan.krakow.pl

Currently I am head of the project of the *Great Dictionary of Polish* (pol.: *Wielki słownik języka polskiego*). The project is coordinated by the Institute of Polish Language at the Polish Academy of Sciences and carried out in collaboration with linguists and lexicographers from several other Polish academic centers. This is an electric, "digitally born" dictionary, encoded as a database, free available on the Internet. This is a typical general dictionary of contemporary Polish. The microstructure of the dictionary contains i.a. definitions, chronology, etymology, inflection (for each entry full inflectional paradigm is given), collocations, quotations, thematic classification, normative information, usage notes. The main source database of the dictionary is the *National Corpus of Polish* (pol. *Narodowy Korpus Języka Polskiego*]. The work is still in progress, by the end of 2013 about 20 000 entries was published.

URL (only Polish version so far): http://wsjp.pl
Description of the dictionary in English:
Żmigrodzki P., 2011: Polish Academy of Sciences Great Dictionary of Polish. History, presence, prospects, „Studies in Polish Linguistics", vol. 6, 2011, s. 7–26. online: http://wsjp.pl/strony_opisowe/WSJP%20studies-6.pdf

Contributor(s):

- **Dan Cristea**, University of Iaşi, Romania, dcristea@info.uaic.ro
- **Marius Clim**, The “Alexandru Philippide” Institute of Romanian Philology, Romania, marius.clim@gmail.com

As promised, here below you can find information about Romanian lexicographic and corpora projects we are aware of:

1. The electronic form of the Thesaurus Dictionary of Romanian Language (eDTLR) - a project financed by Ministry of Research => (mainly retro-) digitisation of the largest dictionary of Romanian, built and printed by Romanian Academy in a period of over one century (1907-2010) => XML TEI-P5 logical form (still having some content and structure bugs) => for the moment, restricted only to the use of researchers.

2. CLRE - a project financed by the Ministry of Research: digitization of 100 Romanian dictionaries (from the 17th century until now, including dictionaries in cyrillic alphabet, latin alphabet and some bilingual dictionaries) and their alignment at entry level. This instrument will facilitate the lexicographic work and also other research concerning language evolution.

3. Studies of diachronic Romanian morphology: how word forms found in the eDTLR citations of word senses can contribute to inferring old morphological paradigms.

4. A research on etymological chains addressing the origins of Romanian: the idea is to build a kind of etymological graphs that show the consecutive formation of words in different languages till their actual form in the Romanian language, also connected with the chronology. If this is to be generalised at an European level, I believe very interesting statistical maps could be drawn, showing how words "migrated" and were transformed at different times, and correlations with historical periods could be inferred.

5. We mention also, a large project involving two Computer Science institutes of the Romanian Academy (In Bucharest and Iasi), started at 1 January 2014, with the aim to build the Representative Corpus of Contemporary Romanian Language, in which textual and speech data will be collected in a balanced manner to represent different genres, domains and styles. Access will be permitted through a query language that will make possible to find words in context. The corpus will be exhaustively annotated (automatically) for sentences, tokens and part of speeches and, partially (manually and automatically), for syntax structure, semantic relations, etc.

Contributor(s):

- **Halldora Jonsdottir**, The Árni Magnússon Institute for Icelandic Studies, Department of Lexicography, Reykjavík, Iceland, halldo@hi.is
- **Thordis Ulfarsdottir**, The Árni Magnússon Institute for Icelandic Studies, Department of Lexicography, Reykjavík, Iceland, disa@hi.is

The ISLEX-project

ISLEX is an on-line multilingual dictionary between modern Icelandic and the Scandinavian target languages (TLs) Danish, Norwegian, Swedish, Faroese and Finnish. The dictionary thus combines six languages in a single database. It is accessible on the web, free of charge.

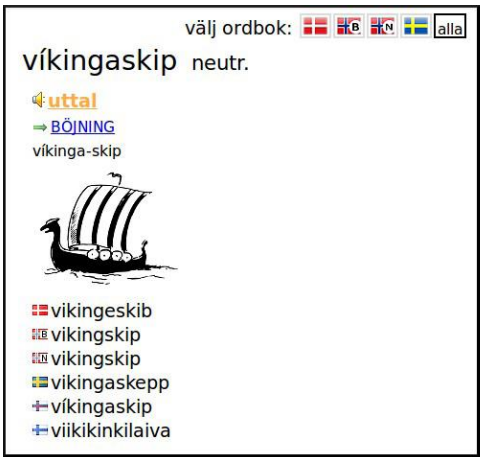
The project is organised in the following way. The largest editorial staff, responsible for the description of the Icelandic source language and for the development and maintenance of the database, is located at *The Árni Magnússon Institute for Icelandic Studies* in Reykjavík, Iceland. Its partners and responsible for the Scandinavian TLs are *The Society for Danish Language and Literature* in Copenhagen (for Danish), *The University of Bergen*, Norway (for both Norwegian language standards: BokmÍl and Nynorsk), *the University of Gothenburg*, Sweden (for Swedish), *The University of the Faroe Islands* (for Faroese) and *Helsinki University* (for Finnish).

The dictionary was opened to the public in November 2011 even if some parts of it were not completed at the time. The work on Icelandic, Danish, Norwegian and Swedish has since been finished, but the work on Faroese and Finnish is still in progress and these dictionaries are not yet open. ISLEX is constantly being updated as needed, e.g. by adding new lemmas and making corrections.

The project has mainly been funded by the governments of the six participating countries. ISLEX runs in a Postgres database and uses Linux Operating System.

ISLEX on the web: www.islex.hi.is (Icelandic), www.islex.dk (Danish), www.islex.no (Norwegian), www.islex.se (Swedish)

Example from the ISLEX-dictionary:



välj ordbok: alla

víkingaskip neutr.

uttal

⇒ BÖJNING

víkinga-skip

vikingeskib
vikingskip
vikingskip
vikingaskepp
víkingaskip
viikikinkilaiva

Contributor(s):

- **Fredrik Norlindh**, Keewords, Sweden, fredrik.norlindh@keewords.com

Keewords

Keewords is an online service that helps people to learn the words of a new language, using the most frequently used words of each language (www.keewords.com). Keewords was one of the founders of the Kelly project (http://www.kellyproject.eu/) which was the mainstay of our language learning material. The project resulted in 72 directed language pairs. Words from that material were lemmatized, part-of-speech tagged, categorized and assigned a difficulty level. Category and difficulty were used to create suitable word lists for users. We have also created similar collections between German, French, Spanish and Korean and Swedish and English.

Users can create their own wordlists using our tool PlayListMaker (https://www.keewords.com/en/playlists/create). User created word lists can be searched and shared. For example teachers can share homework glossary lists with students. Playlistmaker has a type-ahead function and users either pick presented word pairs to have in a word list or alter a pair or create a new one. We want to expand the database as much as possible. We will follow statistics which word pairs users chose and also new word pairs users add themselves. The statistics will be used to show popular word pairs and indicate which types of words users are interested in.

One of many things we are working on right now is expansion of our database with inflections. Later on, we would like to have synonyms, antonyms, phrases, sentence examples, more word pairs and more language pairs.

Contributor(s):

- **María José Domínguez Vázquez**, University of Santiago de Compostela, Spain, majo.dominguez@usc.es

LEXICOGRAPHIC PORTAL: MODULAR MULTILINGUAL ONLINE DICTIONARY BASED ON AN ANNOTATED CORPUS OF THE NOMINAL PHASE.

Management of the project: Prof. Dr. María José Domínguez Vázquez
Members of the project in the European Network of e-Lexicography:
Mónica Mirazo Balsa
Prof. Dr. María Dolores Sánchez Palomino
Prof. Dr. Carlos Valcárcel Riveiro.
Code: FFI2012-32456 (sponsored by Ministerio de Economía y Competitividad, Spain), 2013-2015

SUMMARY

The main goal of this project is the making of an annotated, computerized Spanish – German – Galician – Italian - French corpus and its management systems for the analysis of the noun phrase and its combinations using valential parameters, as well as its implementation in a multilingual lexicographic on-line dictionary. This research is based on the valence theory, in the analysis of vast textual corpora (which provides a solid empirical basis for the lexicographic description) and on previous results of the CSVEA project (INCITE09204074 PR). This new project has a interdisciplinary nature with contributions, not only from different philologies, but also from translation studies as well as from computational and corpus linguistics, providing a vast number of analytic combinations and future applications (monolingual, contrastive and interlingual aspects).

We are thus speaking of an innovative lexicographic portal with plenty of search and information options, based on three central themes: a) multiplicity of languages and a detailed reversible contrastive valential information, b) attention to the on-line architecture of the information, to the diverse user typologies and to the new analytical methods in lexicography- and c) the modular nature of the web portal.

The goals set are: 1. The making of an annotated multilingual corpus and a modular, reversible and multilingual lexicographic on-line portal, based on valential grammar and lexicography, 2. The development of monolingual and reversible contrastive dictionaries for the different considered languages, based on the exploitation of textual corpora as an empirical base for the lexicographic work and on the research on user typology and information architecture, 3. The inclusion of other languages and word classes.

Among the main problems posed by this project, it could be mentioned:

- The design of an adequate database for the multilingual nature of the project, particularly in terms of data management and retrieval.
- The coordination of researchers working on different language and lexicographical traditions.

The collection and management of a high amount of data on a relatively short schedule.

Contributor(s):

- **Maria Tuulik**, The Institute of the Estonian Language, Estonia, maria@eki.ee
- **Jelena Kallas**, The Institute of the Estonian Language, Estonia, jelena.kallas@eki.ee

The Basic Estonian Dictionary: first monolingual L2 learners` dictionary of Estonian

This poster is a report on the lexicographical project completed by the Institute of the Estonian Language by the end of 2013. Dictionary will be published both in print and online in spring 2014.

The Basic Estonian Dictionary (henceforth BED, Kallas, Tuulik 2011) contains several types of lexicographic information: pronunciation, morphological (inflectional) information, definition, word formation, government and collocation patterns, multi-word phrases and semantically related words (synonyms, antonyms).

The pronunciation is presented mostly only online by sound recordings (mp3 audio files) for the most important forms. Morphological (inflectional) information is generated automatically, utilising a rule-based module that has generated the inflected forms as well as the morphonological marking (degree of quantity, etc.) for the BED. The definitions are written using a restricted defining vocabulary of 4,500 words, i.e the list of the headwords of the BED. As the compounding and derivation are the main productive devices for forming new lexical items in Estonian, the word formation information is presented by linking the subentries of the derivated words to the entries of their base forms. As the BED is an active dictionary, the explicit presentation of syntagmatic relations (government and collocational patterns and multi-word phrases) are of the utmost importance. The most frequent government and collocation patterns are analysed and selected using the Sketch Engine corpus query system (Kilgarriff et al. 2004; about Estonian module: Kallas, Tuulik, Jürviste 2012, Kallas 2013). The usage is illustrated by means of example sentences. The semantically related words (synonyms, antonyms) are presented by linking them to each other (in the paper version by the cross-references).

There are ca 400 illustrations in the BED. These are single illustrations with the legend, structural illustrations (particular objects are highlighted by means of arrows), functional illustrations (mostly for adpositions), scenic illustrations (mostly for phrasal verbs), nomenclatory illustarations (e.g. *body*) and enumerating illustrations (e.g. *insect*, *animal*, *flower*, etc.).

The study skills section (16 pages in the paper version) covers several topics, e.g. using numbers, writing informal and business letters and emails, filling in forms, presenting CVs, making phone calls, etc. In the online version the Study Skills section is presented separately on the dictionary interface.

There has been prepared the additional interface of the BED for presenting the dictionary content in sign language using video recordings. For every sign the BED database contains information about the initial hand form, the location where the sign is articulated (i.e. face, lips, cheek, chest, neutral space, etc.) and the movement with which the sign is formed. Based on these three parameters it is possible to search for a certain sign choosing the hand form, the location, and the movement of the sign. This enables the deaf dictionary user to find the Estonian equivalent for a sign.

The BED project is a single database that contains all dictionary related data. XML-based compilation allows us to generate different outputs: for example specialised dictionaries based on partial database output (see Kallas, Langemets 2012). The poster will illustrate also opportunities for reuse of the BED database in order to generate specialized dictionaries (e.g. dictionary of government and collocations).

References

Jürviste, Madis; Kallas, Jelena; Langemets, Margit; Tuulik, Maria; Viks, Ülle (2011). Extending the functions of the EELex dictionary writing system using the example of the Basic Estonian Dictionary. In: *Electronic Lexicography in the 21st Century New Applications for New Users: Proceedings of eLex 2011, Bled, 10-12 November 2011.* (Eds.) Kosem; Iztok; Kosem, Karmen. Ljubljana: Trojina, Institute for Applied Slovenian Studies, 2011, 106–112.

Kallas, Jelena; Langemets, Margit (2012). Automatic Generation of Specialized Dictionaries Using the Dictionary Writing System EELex. In: *Human Language Technologies – The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012. (Toim.) Tavast, Arvi; Muischnek, Kadri; Koit, Mare.* IOS Press, 2012, (Frontiers in Artificial Intelligence and Applications), 103–110.

Kallas, Jelena; Tuulik, Maria (2011). Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. In: *Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics*, Vol. 7, 59–75. DOI: http://dx.doi.org/10.5128/ERYa7.04

Kallas, Jelena; Tuulik, Maria; Jürviste, Madis (2012). Leksikograafilise tarkvara Sketch Engine eesti keele moodul. In: *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEF*, 3-2, 57–77.

Kallas, Jelena (2013). Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. [Syntagmatic relationships of Estonian content words in corpus and pedagogical lexicography]. PhD dissertation. Tallinn: Tallinna Ülikool.

Kilgarriff, A., Rychly, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds.) In: *Proceedings of the XI Euralex International Congress*. Lorient: Université de Bretagne Sud, 105–116.

Contributor(s):

- **Petter Henriksen**, Kunnskapsforlaget, Norway, petter.henriksen@kunnskapsforlaget.no

The Norwegian Academy's Dictionary (Det Norske Akademis Store Ordbok) is the national dictionary for the majority variant of Norwegian *bokmål*. It will be published in 2017, solely on the Internet, at naob.no. Among its features are: export to html from xml format in the editorial database, expandable article parts, hyperlinked references, hyperlinked table of contents for long articles, forum for contact between net users and editorial staff, systematic linking to relevant supplementary material on the net, platform for continuous updating and development in the future. Anticipating the main publication in 2017, we have now started publicizing a selection of new words, coordinated with a weekly contest on one of Norway's more popular radio programs.

Contributor(s):

- **Christine Möhrs**, The Institute for the German Language, Germany, moehrs@ids-mannheim.de

The project I am working for is called *elexiko*. *Elexiko* is a lexicological-lexicographic project based at the *Institut für Deutsche Sprache* (IDS) in Mannheim (cf. Haß 2005, Klosa et al. 2006, Klosa 2011). This dictionary was specifically designed for online publication and is one of the dictionary projects included in the dictionary portal OWID (Online Wortschatz Informationssystem Deutsch). The lexicographers of the project *elexiko* compile a reference work that explains and documents contemporary German. After publishing the complete list of headwords (taken exclusively from the *elexiko*-corpus) in the Internet in 2003, the dictionary was filled with sense independent information for each headword generated automatically or semi-automatically from the underlying corpus (for example details on spelling or syllabification). The next step was the publication of 250 headwords, which were defined as the demonstration module (*Demonstrationswortschatz*) which has been fully lexicographically described. The module *Lexikon zum öffentlichen Sprachgebrauch* (dictionary on public discourse) followed and we are working on this module until now. It contains approximately 2,800 entries selected mainly according to their (high) frequency in the *elexiko*-corpus, such as Europa, Arbeit or Staat.

The technical background enables us to offer the list of headwords and fully lexicographically described headwords in the Internet along with specific search features to users. Dictionary consultants can find single headwords but also can look up groups of lexemes with the same semantic, syntactic, or morphological characteristics (the search options will be extended continually). *Elexiko* offers also auditory examples for selected groups of headwords and illustrations connected to a meaning of a word.

homepage of the project:

http://www.owid.de/wb/elexiko/start.html

lexemes with auditory examples and illustrations:

http://www.owid.de/artikel/249818?module=elex_b

http://www.owid.de/artikel/26859/Bl%C3%BCte?module=elex_b

Haß, Ulrike (Hg.) (2005): Grundfragen der elektronischen Lexikographie. elexiko - das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York. de Gruyter. Schriften des Instituts für Deutsche Sprache.

Klosa, Annette/Schnörch, Ulrich/Storjohann, Petra (2006): ELEXIKO - A Lexical and Lexicological, Corpus-based Hypertext Information System at the Institut für Deutsche Sprache, Mannheim. Alessandria. Edizioni dell'Orso. In: Proceedings of the Twelfth EURALEX International Congress, Torino, Italia, 6 - 9 September 2006. 425-429.

Klosa, Annette (Hg.) (2011): elexiko. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. Tübingen. Gunter Narr. Studien zur Deutschen Sprache. Forschungen des Instituts für Deutsche Sprache.

Contributor(s):

- **Frank Michaelis**, The Institute for the German Language, Germany, michaelis@ids-mannheim.de

In the moment, i'am part of the OWID project team at the Institut für Deutsche Sprache (IDS) in Mannheim.

OWID, the Online-Wortschatz-Informationssystem Deutsch (Online German Lexical Information System) is a lexicographic internet portal for various electronic dictionary resources that are being compiled at the IDS. The main emphasis of OWID is on academic lexicographic resources of contemporary German. The dictionaries included in OWID range from a general monolingual dictionary (elexiko) to a dictionary of neologisms, discourse dictionaries, a dictionary of proverbs and fixed multiword expressions, and a dictionary of German communication verbs (for a complete list see http://www.owid.de/wb/owid/start.html).

OWID is a typical example of a dictionary net (in the sense of (Engelberg & Müller-Spitzer, forthcoming), as it provides inner, outer and external access to the included dictionaries, inter-dictionary cross-references and an integrated layout of portal and individual dictionaries. OWID is a constantly growing resource for academic lexicographic work in the German language.

Contributor(s):

- **Radovan Garabik**, Ľ. Štúr Institute of Linguistics, Slovakia, garabik@kassiopeia.juls.savba.sk

Multilingual glossary based on WordNets

==

The aim of the project is to create a multilingual glossary with Slovak as the central language, predominantly oriented towards non-linguistic audience (e.g. students of Slovak as a foreign language). The project is inspired by the Multilingual Glossary of Synsets (http://metashare.korpus.sk/repository/browse/multilingual-glossary-of-synsets/3c1af05a6ad811e2bd1f00163e0000789a0c01d7093a4af093714bffe431aa71/), it is however not a continuation, but an independent project with different goals and audience. The languages planned to be included in the dictionary are Croatian, Polish, French and English as a pivot language.

The glossary will be built upon existing Slovak language WordNet, which links the synsets to the Princeton WN (v. 3.0), and other languages WordNets. The connections between Slovak and respective languages will be created automatically, using existing mappings to Princeton WN, with manual or semi-manual disambiguation in problematic cases. Since the project includes Slavic languages, special attention has to be paid to perfective/imperfective and reflexive pairs of verbs, where it is mandatory to include correct correspondence between Slovak and the other languages.

The WordNet and inter-language relations will serve as a backend, the dictionary will build upon the database and provide seamless information about semantic equivalents between Slovak and the selected second language. The planned license for the data is CC-BY-SA, Affero GPL and GFDL.

no URL since the project is at its beginning

Collocation dictionary

======================

Electronic dictionary of Slovak collocations is being compiled in collaboration with the University of St. Cyrilus and Methodius in Trnava. Institute in Bratislava. The project on Slovak collocations is the first of its kind in Slovakia and is aimed at the registration and description of multiword lexemes and phrasemes as well as typical collocations with wide collocability. The dictionary provides an overview of the combinatorial behaviour of words, primarily the most frequent nouns extracted from the Slovak National Corpus database. The combinatorial potentials of words are the basis for the creation of so-called collocational templates which are basis for the patterns of collocations. The collocations are extracted from the Slovak National Corpus based on various statistical criteria and manually verified and modified. The final version of the dictionary will not only be available for users electronically but also in a published form.

URL: http://vronk.net/wicol

Slovak valence dictionary

=========================

The main purpose of the project is to create the Slovak version of the Czech PDT-Vallex valency lexicon which has been created as a part of the Prague Dependency Treebank (PDT). The first step consists of automatic translation of PDT-Valex into Slovak and automatic creation of Slovak language valency frames. Subsequently we plan to analyse the accuracy of translation (since Slovak is very close to Czech, there will not supposedly be a marked difference in the valency frames of many verbs; however there are notable differences for often used ones), using the data from Slovak National Corpus and Syntactically annotated corpus of Slovak (which is compatible with PDT, with only very minor differences). The second step entails creation of tools for semi-automatic validation of selected aspects of valency frames, together with manual proofreading of most frequent verbs. The goal is to obtain valency lexicon consisting of a small manually proofread core set of verbs and with much bigger database of Slovak verbs, which will nevertheless be of reasonably high accuracy. The resulting dictionary might be further used for contrastive analysis of syntactic and semantic properties of both languages. The lexicon will be used in creation the Slovak verbal collocation lexicon a database of verbal multiword expressions, which will be the third lexicon of this type (Noun and Adjective collocation lexicons) compiled at the Slovak National Corpus Department. Valency properties of verbs is reflected in their collocational potentiality which enables words to co-occur in fixed syntactic relations; some of them form fixed multiword expressions. Valency frames will be used in creating collocational structures (collocations) of Slovak verbs. The lexicon could be a helpful lexical resource and also used in automatic verification of collocation dictionary of verbs.

no URL since the project is at its beginning

Database of nouns

=================

The aim of the database is to compile a representative database of Slovak language nouns with complete inflectional paradigms, corpus examples and frequencies for non-linguistically oriented audience. For each noun, the database contains its inflected forms, corpus example for each combination of grammatical case and number and an intuitive visual representation of the absolute frequency of the lexeme, and the relative frequency of each wordform.

Further plans include expanding the database to cover other (inflected) parts of speech, namely verbs, adjectives and participles.

preliminary alpha version of the database is accessible at:
http://labs.juls.savba.sk/?d=noundb

Contributor(s):

- **Tarja Heinonen**, Institute for the Languages of Finland, Finland, tarja.heinonen@kotus.fi
- **Riina Klemettinen**, Institute for the Languages of Finland, Finland, riina.klemettinen@kotus.fi

Kielitoimiston sanakirja – The New Dictionary of Modern Finnish

The New Dictionary of Modern Finnish is a monolingual general-purpose dictionary containing about 100,000 entries. The dictionary is revised and updated on a continual basis, and it is available both in print and in electronic form. The latest editions came out 2012. The digital edition of the dictionary offers the inflectional paradigm for each inflecting headword and allows searches by several criteria. It also includes a separate glossary of 21,000 Finnish place names and their inflected forms.

One of our future challenges is to add new features to the electronic dictionary. Currently we are tagging idiomatic examples in order to make them more easily searchable. We also plan to add linking to usage information (prescriptive notes, grammatical clarifications and the like) on specific entries. One of the goals further in the future is to develop a dictionary portal involving all the dictionaries of the Institute for the Languages of Finland.

Contributor(s):

- **Egon W. Stemle**, EURAC, Italy, egon.stemle@eurac.edu

I am Cognitive Scientist, this means I study skills like perception, thinking, learning, motor function, and language by combining the humanistic and analytical methods of the arts and the formal sciences. I am working on computer aided fabrication of ontologies from large document repositories, the technological feasibility thereof and the utilization of cross-linked structured data in applications, and on tools for editing, processing, and annotating linguistic data. My driving force is the question why humans handle incomplete and – more often than not – inconsistent structured-concepts just fine, whereas computational processes are often of little avail or fail completely.

Currently, i am working in the DiDi project where we analyse the linguistic strategies employed by users of social network sites (SNS). The data analysis will focus on South Tyrolean users and we will investigate how they communicate with each other. In regions of the German speaking area where dialect is frequently used in different communicative contexts, regional and social codes are often also used in written communication and in computer mediated communication. Another interesting but more general aspect of the new media is connected to the emerging linguistic and social practices (new literacy) that are usually discussed at the interface of linguistic and social description. One of the main research questions in DiDi is whether people of different age use language on SNS in a similar way or in an age-specific manner.

Contributor(s):

- **Anna Dziemianko**, Adam Mickiewicz University, Poland, danna@wa.amu.edu.pl

The projects I am currently involved in concern:

- optimal definition formats in (e)dictionaries,
- ordering senses in (e)dictionaries,
- colors in e-dictionaries,
- optimizing the presentation of collocations in e-dictionaries
- information retrieval from polysemous entries

Contributor(s):

- **Carole Tiberius & Tanneke Schoonheim**, Instituut voor Nederlandse Lexicologie (INL), {carole.tiberius,tanneke.schoonheim}@inl.nl

The *Algemeen Nederlands Woordenboek* (ANW; Dictionary of Contemporary Dutch) is a corpus-based, scholarly dictionary of contemporary standard Dutch in the Netherlands and in Flanders, describing the Dutch vocabulary from 1970 onwards.
The ANW can be characterised as an online dictionary under construction. The project started in 2001 and a first version of the full dictionary will be completed at the end of 2018. In December 2009 a demo version was launched and new dictionary articles are being added on a regular basis (with an average of 4 updates per year).[1]

The ANW is a corpus-based dictionary and the ANW corpus was compiled specifically for the project and consists of several subcorpora: literary texts (20%), newspaper material (40%), domain-specific texts (35%), and a sub corpus with texts containing neologisms (5%). Originally, corpus compilation was completed in 2004, when a corpus size of just over 100 million tokens[2] was reached. However, we are now in the process of moving from a static corpus to a dynamic corpus by adding new material to the various subcorpora starting with newspaper material.

The dictionary provides information on the core vocabulary of contemporary Dutch and it pays special attention to neologisms. These are collected and examined by the editorial staff, who especially pay attention to their relation in form and/or meaning to other Dutch words. Each entry contains information on spelling, spelling variation, inflection, part of speech and morphology. For etymology links are provided to the etymological database of Dutch (www.etymologiebank.nl). The meaning of the entries is mostly explained by analytical or morpho-semantic definitions, combined with synonyms and antonyms. On this level there are links to corresponding entries in the historical dictionaries of Dutch (http://gtb.inl.nl). Regional variation is given, enabling the user to distinguish Dutch from the Netherlands, Dutch form Belgium and Dutch from Suriname. Frequently used combinations are given, mostly provided by word sketches from the Sketch Engine (Kilgarriff et al. 2004), as well as collocations and proverbs. A special feature of the ANW entry is the so-called semagram, in which the meaning of the word is described in terms of slots and fillers, thus allowing the lexicographer to provide the user with more and more accurate information on the word than could be given in any definition. For more information on the semagram, see Moerdijk 2008. Meaning, combinations, collocations and proverbs are provided with illustrating quotations from the corpus, thus enabling the user to verify the lexicographer's interpretation. An important element of the meaning section is also word family, in which derivations and compounds of the word in focus are given, classified at their actual meaning.

An important feature of the online dictionary is that it offers a range of search strategies, from text understanding to text production. Four search options are distinguished:

a) Word → Meaning, i.e. search for information about a word or phrase;

[1] For more information on the ANW see Schoonheim and Tempelaars (2010) and references on the ANW website: http://anw.inl.nl/show?page=help_publicaties.

[2] A corpus of one hundred million tokens is considered to be large enough for describing the normal use of a language (cf. Hanks 2002: 157).

b) Meaning → Word, i.e. search for a word starting from the meaning;
c) Features → **Words**, i.e. search for words with one or more shared features.
d) Examples, i.e. search for example sentences.

URL: http://anw.inl.nl

Contributor(s):

- **Simon Krek**, Jožef Stefan Institute, Slovenia, simon.krek@guest.arnes.si
- **Iztok Kosem**, Trojina, Slovenia, iztok.kosem@trojina.si

SLOVENE LEXICAL DATABASE

Slovene Lexical Database was created between 2008 and 2012 and represents a comprehensive **syntactic and semantic description** of a selected set of Slovene words. The description was based exclusively on the analysis of reference corpora of Slovene. The wordlist in the Lexical Database was selected from 5,000 most frequent word in FidaPLUS and Gigafida corpora. In addition, we also considered a selection of words from school books in order to accomodate the needs of school population.

The purpose of creating Slovene Lexical Database is, first, to fill the existing gap in comprehensive lexical description of Slovene both from the point of view of detecting changes in the modern vocabulary of Slovene and of introducing modern lexicographic procedures in Slovene lexicography. It offers information about the meaning of words, their tipical context, stylistic, pragmatic and other peculiarities, fixed expressions and phraseology to general users, school population and learners of Slovene as a foreign language. All information intended for "human" users are worded in the manner known to the user from everyday communication. And secondly, Slovene Lexical Database is designed to provide language data in the form useful for natural language processing applications and language technology tools for Slovene.

The database contains 2,500 entries with 10,946 lexical units: senses, sub-senses, multi-word units and phraseological units.

The concept brings a **new style of semantic description** of Slovene vocabulary which is focused on typical context and based on the image of Slovene as found in real texts.

The database is structured as a network of interrelated semantic and syntactic information about a particular word. **Semantic level** represents the top level in the hierarchy with the lexical unit as its core element. This includes all senses of the headwrd, multi-word expressions and phraseological units. Each sense is described with a short **semantic indicator** and/or **whole-sentence definition** which includes typical syntactic environment of the headword with the relevant number, form and semantic types in a valency frame (**semantic frame**). These are also reflected in a number of **syntactic structures** and corresponding **collocations**. All the higher types of information are confirmed by a selection of **corpus examples**.

Multi-word expressions and **phraseological units** are treated independently from particular senses of the headword and have their own internal structure which requires the same types of information as single-word entries or senses.

Web site (downloadable): http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza